

1 WHAT IS CLAIMED IS:

1 1. In a data processing system, a method for creating a database from
2 information found on a plurality of web pages using a first classifier and a second
3 classifier, said method comprising:

4 a) defining first regularities and second regularities, said first regularities
5 being patterns which are expected to be found in information in said web pages, and said
6 second regularities being patterns which are not expected to be found in all said web
7 pages;

8 b) initially providing descriptions of said first regularities to a working
9 database; thereafter

10 c) training a first classifier in said working database using said first
11 regularities;

12 d) identifying a candidate subset of the web pages expected to have said
13 second regularities; thereafter

14 e) tentatively identifying and tagging, in said candidate subset of the web
15 pages, elements having said first regularities, by using said first classifier to obtain first
16 tentative labels;

17 f) training a second classifier using said first tentative labels;

18 g) tentatively identifying elements having specific combinations of said
19 first regularities and said second regularities using said first classifier and said second
20 classifier to obtain second tentative labels for said elements of said candidate subset;
21 thereafter

22 h) outputting, said second tentative labels as permanent labels associated
23 with said elements of said candidate subset of web pages.

1 2. The method of claim 1 further including:

2 h) deciding whether to retrain said second classifier with said second
3 tentative labels.

1 3. The method according to claim 2, further including:

2 f) training the second classifier using said second tentative labels.

1 4. The method according to claim 2, further including

2 g) collecting said permanent labels associated with said elements of said
3 candidate subset of web pages;

4 h) retraining said first classifier in response to said permanent labels.

1 5. The method according to claim 1 wherein said second classifier
2 treats selected first regularities differently than said first classifier treats said first
3 regularities such that said second regularities contradict said first regularities.

1 6. The method according to claim 5 wherein said outputting step
2 further includes ignoring training results of said first classifier.

1 7. The method according to claim 5 wherein said outputting step
2 further includes combining training results of said first classifier and said second
3 classifier.

1 8. In a data processing system, a method for learning and combining
2 global regularities and local regularities for information extraction and classification, said
3 method comprising the steps of

4 a) initially providing descriptions of said global regularities to a working
5 database, said global regularities being patterns which may be found over an entire
6 dataset; thereafter

7 b) identifying a candidate subset of the dataset in which local regularities
8 may be found; thereafter

9 c) tentatively identifying elements having said global regularities in said
10 candidate subset to obtain first tentative labels, said first tentative labels being useful for
11 tagging information having identifiable similarities; thereafter

12 d) attaching said first tentative labels onto said identified elements of said
13 candidate subset; thereafter

14 e) employing said attached first tentative labels via one of a class of
15 inductive operations to formulate first local regularities; thereafter

16 f) tentatively identifying elements having specific combinations of said
17 global regularities and said local regularities to obtain attached second tentative labels;
18 thereafter

19 g) testing if estimated error rate is within a preselected tolerance or if a
20 steady state in said attached second tentative labels is evident; and if true, then rating

21 confidence of said attached second tentative labels and converting selected ones of said
 22 attached second tentative labels to confidence labels upon achieving a preselected
 23 confidence level and then outputting data with said confidence labels; otherwise
 24 h) employing said second tentative labels via said operation on said
 25 candidate subset to formulate second local regularities, and
 26 i) repeating from step f) until said confidence labels have been developed.

1 9. The method according to claim 8 wherein said initial global
 2 regularity providing step comprises manually inputting descriptions of said global
 3 regularities.

1 10. The method according to claim 8 wherein said initial global
 2 regularity providing step comprises obtaining said global regularities from a further one
 3 of said class of said inductive operations applied to a subset of said dataset, said subset of
 4 said dataset having been manually labeled.

1 11. The method according to claim 10 further including developing
 2 refined global regularities comprising the steps of:

3 l) collecting confidence labels from at least one of said candidate subsets to
 4 obtain global confidence labels;

5 m) employing said global confidence labels on candidate subsets along
 6 with said manually labeled dataset via one of said class of inductive operations to
 7 formulate said refined global regularities;

8 n) providing descriptions of said refined global regularities to said working
 9 database; thereafter

10 o) identifying a next candidate subset of the dataset in which local
 11 regularities may be found; thereafter

12 p) tentatively identifying elements having said refined global regularities
 13 in said candidate subset to obtain next tentative labels; thereafter

14 q) attaching said next tentative labels onto said identified elements of said
 15 next candidate subset; thereafter

16 r) employing said attached next tentative labels via one of the class of
 17 inductive operations to formulate next local regularities; thereafter

18 s) tentatively identifying elements having specific combinations of said
19 refined global regularities and said next local regularities to obtain attached next second
20 tentative labels; thereafter
21 t) testing if estimated error rate is within a preselected tolerance or if a
22 steady state in said next second tentative labels is evident; and if true, then rating
23 confidence of said attached next second tentative labels and converting selected ones of
24 said attached next second tentative labels to confidence labels upon achieving a
25 preselected confidence level and then outputting data with said confidence labels;
26 u) otherwise employing said next second tentative labels via said operation
27 on said candidate subset to formulate next second local regularities, and
28 v) repeating from step s).

1 12. The method according to claim 11 further including the steps of:
2 applying said data with confidence labels to further subsets of said dataset
3 to investigate further subsets for local regularities.

1 13. The method according to claim 8 further including developing
2 refined global regularities comprising the steps of:
3 l) collecting confidence labels from at least one of said candidate subsets to
4 obtain global confidence labels;
5 m) employing said global confidence labels on candidate subsets via one
6 of said class of inductive operations to formulate said refined global regularities;
7 n) providing descriptions of said refined global regularities to said working
8 database; thereafter
9 o) identifying a next candidate subset of the dataset in which local
10 regularities may be found; thereafter
11 p) tentatively identifying elements having said refined global regularities
12 in said candidate subset to obtain next tentative labels; thereafter
13 q) attaching said next tentative labels onto said identified elements of said
14 next candidate subset; thereafter
15 r) employing said attached next tentative labels via one of the class of
16 inductive operations to formulate next local regularities; thereafter

17 s) tentatively identifying elements having specific combinations of said
18 refined global regularities and said next local regularities to obtain attached next second
19 tentative labels; thereafter
20 t) testing if estimated error rate is within a preselected tolerance or if a
21 steady state in said next second tentative labels is evident; and if true, then rating
22 confidence of said attached next second tentative labels and converting selected ones of
23 said attached next second tentative labels to confidence labels upon achieving a
24 preselected confidence level and then outputting data with said confidence labels;
25 u) otherwise employing said next second tentative labels via said operation
26 on said candidate subset to formulate next second local regularities; and
27 v) repeating from step s).

1 14. The method according to claim 13 further including the steps of:
2 applying said data with confidence levels to further subsets of said dataset
3 to investigate further subsets for local regularities.

1 15. In a data processing system, a method for learning and combining
2 regularities of a first level and regularities of at least a second level and a third level for
3 information extraction and classification, said method comprising the steps of:
4 a) determining a hierarchy of levels from most global level to most specific
5 level;
6 b) beginning at the most global level, training a classifier at the selected
7 level by initially providing descriptions of regularities at said selected level to a working
8 database, said selected level regularities being patterns which may be found over a
9 selected portion of a selected dataset; thereafter
10 c) identifying a candidate subset of the selected dataset in which next more
11 specific regularities may be found; thereafter
12 d) tentatively identifying elements having said selected regularities in said
13 candidate subset to obtain first tentative labels, said first tentative labels being useful for
14 tagging like information; thereafter
15 e) attaching said first tentative labels onto said identified elements of said
16 candidate subset; thereafter
17 f) employing said attached first tentative labels via one of a class of
18 inductive operations to formulate first local regularities; thereafter

19 g) tentatively identifying elements having specific combinations of said
20 global regularities and said local regularities to obtain attached second tentative labels;
21 thereafter
22 h) testing if estimated error rate is within a preselected tolerance or if a
23 steady state in said attached second tentative labels is evident; and if true, then rating
24 confidence of said attached second tentative labels and converting selected ones of said
25 attached second tentative labels to confidence labels upon achieving a preselected
26 confidence level and then outputting data with said confidence labels;
27 i) otherwise employing said second tentative labels via said operation on
28 said candidate subset to formulate second more specific regularities,
29 j) repeating from step g); and
30 k) repeating from Step b) for each successive more selective level of
31 regularity.